Вестник Тюменского государственного университета.
2014. № 7. Физико-математические науки. Информатика. 242-250

© Е.Г. БРУНОВА, Ю.В. БИДУЛЯ
Тюменский государственный университет
egbrunova@mail.ru, bidulya@yandex.ru

УДК 81'322

АЛГОРИТМ С ЭЛЕМЕНТАМИ ФОРМАЛЬНОЙ ГРАММАТИКИ ДЛЯ КОНТЕНТ-АНАЛИЗА МНЕНИЙ

ALGORITHM WITH FORMAL GRAMMAR ELEMENTS FOR SENTIMENT ANALYSIS

АННОТАЦИЯ. Исследование, выполненное в области математической лингвистики, посвящено анализу субъективной информации, содержащейся в пользовательском контенте. Составлен оценочный лексикон (583 единицы), специализированный по предметной области (банковское дело) и языку (русский). В оценочный лексикон включены следующие классы слов: положительная лексика, отрицательная лексика, модификаторы, антимодификаторы и инкременты полярности. Представлен алгоритм REGEX с элементами формальной грамматики для контент-анализа мнений. Введены 11 правил формальной грамматики и соответствующие синтаксические модели, которые являются своего рода регулярными выражениями, позволяющими обнаружить определенные элементы текста, упростить каждое предложение и представить текст в целом как формальную модель. На основе предлагаемого алгоритма разработана система SENTIMENTO для оценки качества банковского обслуживания, реализованная в виде интернет-приложения с интерфейсом для апробации модели и ее корректировки. Эффективность предлагаемого алгоритма сопоставлена с эффективностью наивного Байесовского классификатора, в качестве критерия применена мера Ван Ризбергена. Апробация системы на материалах отзывов, опубликованных в народном рейтинге банков на сайте www.banki.ru, показала преимущество разработанного алгоритма. Для одного и того же набора отзывов при использовании предложенного в работе метода величина показателя F1 составила 0.920, в то время как для наивного Байесовского классификатора величина F1 оказалась равна 0.872.

SUMMARY. This study carried out within computational linguistics presents the analysis of the subjective information from user-generated content. The sentiment lexicon (583 items) which is domain-specific (banking) and language-specific (Russian) is built. The sentiment lexicon includes the following classes: positive vocabulary, negative vocabulary, polarity modifiers, anti-modifiers, and increments. The REGEX algorithm with formal grammar elements is proposed. 11 formal grammar rules and the corresponding syntactic models are introduced; they are similar to regular expressions which detect certain text elements, simplify each sentence, and present the text as a formal model. The SENTIMENTO system for evaluating bank service quality is implemented as

an Internet application with an interface for the model testing and its adjustment. The efficiency of the proposed algorithm is evaluated in comparison with the efficiency of the Nanve Bayes Classifier, F1 measure is used as the criterion. The system is tested on the reviews published in the clients' bank rating (www.banki.ru) and the advantage of the proposed algorithm is demonstrated. For the same set of reviews, the F1 value is 0.920 when the proposed method is applied, while it is 0.872 for the Nanve Bayes Classifier.

КЛЮЧЕВЫЕ СЛОВА. Обработка естественного языка, алгоритм, контент-анализ мнений, формальная грамматика, наивный Байесовский классификатор, пользовательский контент.

KEY WORDS. Natural language processing, algorithm, sentiment analysis, formal grammar, Naive Bayes Classifier, user-generated content.

Введение. Исследование, выполненное в области математической лингвистики, посвящено анализу субъективной информации, содержащейся в пользовательском контенте (отзывах потребителей о качестве банковского обслуживания, опубликованных в сети Интернет).

Информацию, содержащуюся в текстах на естественном языке, можно условно отнести к одному из двух типов: факты и мнения. Факты — это объективная информация, описывающая сущности и события, а также их свойства. Мнения - это субъективная информация, описывающая оценку (одобрение или неодобрение) и эмоции человека по отношению к сущностям и событиям, а также их свойствам. Успешные разработки в области поиска, извлечения и обработки информации из текстов на естественном языке, как правило, сосредоточены на задачах, связанных с фактами, и только в последнее десятилетие стали появляться исследования, посвященные поиску, извлечению и обработке мнений [1-5]. Такие исследования в подавляющем большинстве посвящены анализу текстов на английском языке. Что касается публикаций по поводу исследования мнений в русскоязычных текстах, то они только начинают появляться и носят в основном обзорный характер [6-9].

В настоящее время анализ субъективной информации является одной из наиболее перспективных задач в области математической лингвистики. Всплеск интереса к данной проблеме обусловлен развитием социальных сетей, блог-платформ и других технологий, с помощью которых возникает большой объем пользовательского текста, и, следовательно, необходимость его изучения в научных и коммерческих целях. Так, производители товаров и поставщики услуг заинтересованы в получении обработанной информации о настроениях потребителей. Потребителей, в свою очередь, при выборе товара или услуги, интересуют мнения других людей, основанные на их личном опыте.

Одним из методов, широко применяемых в контент-анализе мнений, является наивный Байесовский классификатор, основанный на вероятностной модели [10]. Данный классификатор рассматривает текст отзыва как набор слов. Именно простота является его главным достоинством, поскольку она позволяет отказаться от использования таких трудоемких инструментов, как синтаксический анализ, что особенно ощутимо для текстов на русском языке, поскольку «для русского языка до сих пор не решены задачи синтаксического анализа и разрешения анафорических связей, что в значительной мере осложняет более тонкий анализ» [6; 88]. В то же время простота является и главным недостатком

наивного Байесовского классификатора. Текст на естественном языке не сводится к набору слов (как, например, набор ключевых слов статьи или набор слов поискового запроса), а отношения слов на уровне предложения и на уровне целого текста могут существенно влиять на смысл текста и, следовательно, на его оценку.

Целью нашего исследования является разработка улучшенного классификатора, который использует преимущества наивного Байесовского классификатора и минимизирует его недостатки.

Наша гипотеза заключается в следующем: для естественного языка характерно большое разнообразие средств выражения мнений и эмоций, а его представление в виде набора слов не позволяет получить надежных результатов контент-анализа мнений. Вместо трудоемкого синтаксического анализа мы предлагаем набор правил формальной грамматики, подобных регулярным выражениям.

Предлагаемый нами улучшенный классификатор состоит из трех компонентов: 1) тренировочная выборка текстов, в которой анализ мнений выполняется вручную; 2) оценочный лексикон, содержащий подмножества слов, отнесенных к определенному классу; 3) набор формальных правил количественной оценки полярности мнений.

Материал и методы исследования. Для тренировочной выборки случайным образом было отобрано 20 документов — отзывов о качестве банковского обслуживания — с сайта www.banki.ru, народный рейтинг банков (10 положительных и 10 отрицательных). Из данных документов вручную был составлен базовый оценочный лексикон, включающий следующие классы слов: 1) положительная лексика (*доброжелательность, доверие, защита, бесплатный, вежливый, грамотный, благодарить* и т.п.); 2) отрицательная лексика (*безвыходный, бюрократичный, грубый, досадный, конфликт, мрак, нервотрепка, обида, очередь, заблокировать* и т.п.); 3) модификаторы полярности (*не*, *нет, без*); 4) антимодификаторы полярности (*такой, так, настолько*); 5) инкременты полярности (*очень, крайне, самый, единый, никогда, нигде* и т.п.).

Оценочный лексикон формировался следующим образом:

1) из тренировочной выборки вручную были составлены лексиконы модификаторов, антимодификаторов и инкрементов полярности, а также базовые положительный и отрицательный лексиконы, всего 100 единиц. Единицей лексикона в англоязычных разработках является лемма (слово в своей основной форме). Применительно к русскому языку мы использовали усеченное слово, полученное при помощи процедуры стемминга, т.е. основу без окончаний, например, *хорош* вместо *хороший* и *хорошо*.

2) базовые лексиконы были расширены за счет синонимов и антонимов, например, *хорош(ий)→превосходн(ый), отличн(ый), замечательн(ый), плох(ой), нехорош(ий)* и т.д. Для этой цели использовались словари синонимов и антонимов.

3) базовые лексиконы были расширены с помощью поисковых запросов с булевыми операторами по методике В. Хацивассилоглоу и К. МакКьюена [11] (поисковая система *www.google.com* с ограничением поиска по сайту *www.banki.ru*).

После расширения объем оценочного лексикона составил 583 единицы. Оценочный лексикон специализирован по предметной области (банковское дело) и языку (русский).

Результаты и их обсуждение. Для получения численной оценки контент-анализа мнений были разработаны правила двух типов: 1) правила замены слов метасимволами для получения схемы разметки текста; 2) правила обнаружения и подсчета признаков, представленных в виде определенных синтаксических конструкций.

Правила замены описывают последовательное преобразование слов, пробелов и знаков препинания каждого предложения на соответствующие метасимволы. Заменяя текст схемой разметки, составленной из метасимволов, мы сводим аналитическую работу над неструктурированным текстом к использованию хорошо отработанного механизма регулярных выражений.

Ниже показан алгоритм применения правил замены в виде действий, каждое из которых приводит к появлению в схеме разметки группы метасимволов. На вход алгоритма подается неформализованный текст отзыва, представляющий некоторую последовательность символов.

Алгоритм REGEX.

Шаг 1. Разбить текст на предложения, пометив начало и конец каждого предложения метасимволами <S>, </S>, <!/S>, <?/S>, <?!/S>.

Шаг 2. Каждое предложение разбить на части, заменив знаки препинания в середине предложения, кавычки, союзы *и* и *или* на метасимволы <Z>,<Q>, & соответственно.

Шаг 3. Разбить по пробелам на слова.

Шаг 4. Подсчитать количество слов в тексте.

Шаг 5. Пометить слова, набранные прописными буквами, метасимволом CAPS.

Шаг 6. Заменить все прописные буквы строчными.

Шаг 7. Произвести стемминг всех слов.

Шаг 8. Заменить слова из лексиконов на соответствующие метасимволы. Если слово не найдено ни в одном из лексиконов, оно заменяется символом «*».

Шаг 9. Заменить все последовательности «* *» на «*»

Пример получения разметки представлен на рис. 1. Как видно из примера, текст отзыва на выходе представляет строку, состоящую исключительно из метасимволов разметки.

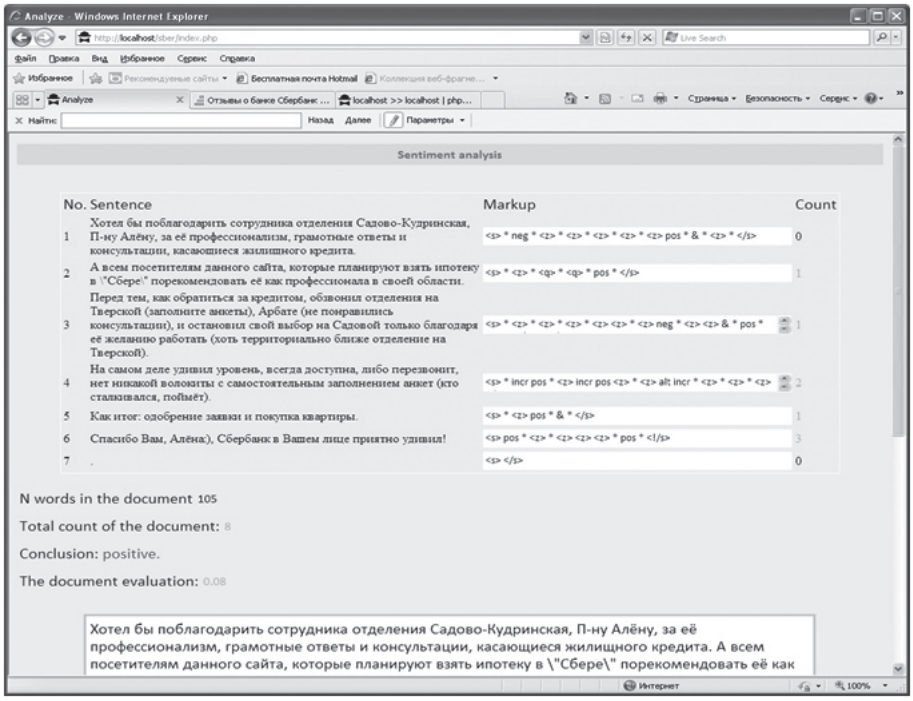

Рис. 1. Вид программного модуля для контент-анализа мнений.

<S> — начало предложения; </S> — конец предложения с точкой; <!/S> — конец предложения с восклицательным знаком; <Z> — знак препинания в середине предложения, & — союз *и/ или*; POS — слово из положительного лексикона; NEG — слово из отрицательного лексикона; ALT — модификатор полярности; W — слово *без*; * — 0 или более любых слов, не входящих в оценочный лексикон.

Правила контент-анализа мнений первоначально были получены в неформализованном виде путем анализа вручную отзывов тренировочной выборки. Далее каждое правило было формализовано с применением метасимволов. Всего было разработано 11 правил. Приведем одно из них.

Правило 1: *Если в промежутке от начала предложения или знака препинания или союза И / ИЛИ, или до следующего знака препинания или союза И / ИЛИ имеется модификатор полярности, то полярность всех слов, входящих в оценочный лексикон, в данном промежутке изменяется на противоположную. Порядок указанных элементов (модификатор, слово с полярностью, любое другое слово) значения не имеет.*

В формализованном виде правило выглядит следующим образом:

<S>|<Z>|& {ALT, *, Any POS} <Z>|&|</S>|<!/S>|<?/S>|<?!/S> →

→ <S>|<Z>* Any NEG *<Z>|&|</S>|<!/S>|<?/S>|<?!/S> →

→ nNEG → -n

где {A, B, C} — группа элементов, которые могут следовать в данном промежутке в любом порядке; Any — любое число элементов; | — разделитель в равной степени допустимых элементов; nNEG — количество слов отрицательного лексикона.

Алгоритм преобразования схемы разметки включает последовательное применение правил замены в соответствии с приоритетами, назначенными в результате экспериментов над текстами отзывов из тренировочной выборки. Например, применение правила 1 приводит к следующему преобразованию схемы разметки (выделено жирным шрифтом):

Платежи проходят очень быстро, деньги ***не зависают****.*

<S>*POS <Z> * ALT NEG </S>→ <S>*POS <Z> * POS </S> →

→ 2POS → +2

На определенном шаге алгоритма подсчитывается количество метасимволов POS и NEG, после чего определяется «сырая» численная оценка полярности мнения каждого предложения. Пример подсчета:

Очередей нет, все чистенько, удобно.

<S>NEG ALT <Z>*POS <Z>POS </S> →

→ <S>POS <Z>*POS <Z>POS </S> → 3POS → +3

Предусмотрена группа правил, направленных на коррекцию «сырой» оценки. Например, правило 10: *если в предложении имеются слова из прописных букв*, а счет предложения положительный, то каждое слово из прописных букв приравнивается к одному слову с положительной полярностью.* В формализованном виде правило выглядит так:

<S> * CAPS * </S> | <!/S> n>0→ n +1

Пример применения:

Хочу поблагодарить за такой сервис и соответственно оценить — ОТЛИЧНО.

<S> *POS*POS* POS/CAPS </S> → 3POS → +3(>0) +1= +4

Итогом работы алгоритма является подсчет общей оценки по всем предложениям отзыва, а также нормализованная оценка, приведенная к числу слов в отзыве.

Предлагаемая модель была апробирована в программном комплексе SENTIMENTO, реализованном в виде интернет-приложения на базе web-сервера Apache. Приложение состоит из двух модулей: модуля администрирования и модуля контент-анализа мнений.

В модуле администрирования предоставляется интерфейс для заполнения оценочного лексикона, при этом возможен как импорт словарей из файла Excel, так и добавление отдельных слов. При внесении слова в базу данных производится его проверка на наличие или отсутствие окончания, так как в оценочном лексиконе хранится только неизменяемая часть слова. Если окончание присутствует, программа производит стемминг слова и предъявляет пользователю результат для проверки и подтверждения. Кроме того, в обязательном порядке слову назначается класс, определяющий то подмножество лексикона, к которому это слово будет отнесено: отрицательная лексика, положительная лексика, модификатор и т.д.

Модуль контент-анализа мнений позволяет ввести текст на естественном языке в поле ввода и нажать на кнопку «Do sentiment analysis» («Выполнить

* *Слово, набранное прописными буквами, в интернет-коммуникации приравнивается к крику, т.е. выражает сильную эмоцию, и, следовательно, должно учитываться при контент-анализе мнений.*

контент-анализ мнений»). Программный код скрипта читает текст из поля ввода и проводит над ним последовательные преобразования в соответствии с описанными выше правилами, вычисляет и выводит на экран текст отзыва, численную оценку, а также заключение системы об отзыве (положительный или отрицательный).

Пользователь имеет возможность подтвердить или опровергнуть оценку системы. Для этого на экране с результатами оценки системы выводится надпись: «Your conclusion» («Ваше заключение») и две кнопки: «Positive» («Положительный») и «Negative» («Отрицательный»). После нажатия на одну из этих кнопок система проверяет, совпадают ли оценки системы и пользователя. Указанные данные используются для сопоставления оценок системы и человека и вычисления значений точности, полноты и меры F1 (меры Ван Ризбергена), вычисляемой по формуле:

$$F1 = \left[\frac{2P * R}{P + R} \right]$$

где P — точность (отношение количества найденных релевантных документов к общему количеству документов, найденных системой), R — полнота (отношение количества найденных релевантных документов к общему количеству релевантных документов) [12; 170].

В программном комплексе реализована возможность мониторинга результатов с целью отладки алгоритма. При вводе текста отзыва пользователю предлагается опция «Show mark-up» («Показать разметку») для просмотра схемы разметки каждого предложения и его численной оценки. На рис. 1 показан вид программы при выводе результатов оценки в режиме мониторинга.

Выводы. Для определения эффективности работы алгоритма REGEX был проведен его сравнительный анализ с наивным Байесовским классификатором. Обучение производилось на корпусе из 50 отзывов, (28 положительных и 22 отрицательных). Далее проводилась тестовая оценка 20 произвольных отзывов.

Эксперимент показал, что для одного и того же набора отзывов при использовании предложенного в работе метода величина показателя F1 составила 0.920, в то время как для наивного Байесовского классификатора величина F1 оказалась равна 0.872. Применение предложенных нами формальных правил позволяет решить одну из задач математической лингвистики — существенно улучшить качество контент-анализа мнений при оценке качества банковского обслуживания. Совершенствование правил и использование методики составления лексикона позволит распространить применение улучшенного классификатора и на другие предметные области.

СПИСОК ЛИТЕРАТУРЫ

1. Carenini, G., et al. Extracting Knowledge from Evaluative Text // Proceedings of the 3rd International Conference on Knowledge Capture. 2005. Pp. 11-18.

2. Hu, M., Liu, B. Mining and Summarizing Customer Reviews // Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2004. Pp. 168-177.

3. Nasukawa, T., Yi, J. Sentiment Analysis: Capturing Favorability Using Natural Language Processing // Proceedings of the 2nd International Conference on Knowledge Capture. Florida, 2003. Pp. 70-77.

4. Pang, B., Lee, L. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarisation Based on Minimum Cuts // Proceedings of the ACL, 2004, Pp. 271-278.

5. Turney, P. Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews // Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. 2002. Pp. 417-424.

6. Ермаков С.А., Ермакова Л.М. Методы оценки эмоциональной окраски текста // Вестник Пермского университета. Вып. 1(19). 2012. С. 85-89.

7. Лукашевич Н.В., Четверкин И.И. Извлечение и использование оценочных слов в задаче классификации отзывов на три класса // Вычислительные методы и программирование. 2011. Т. 12. С. 73-81.

8. Оробинская Е.А., Кочуева З.А. Технологии Text Mining: Обзор методов и задач обработки смысловой информации // Вестник Херсонского национального технического университета. № 2 (38). 2010. С. 348-353.

9. Пазельская А.Г., Соловьев А.Н. Метод определения эмоций в текстах на русском языке // Компьютерная лингвистика и интеллектуальные технологии: «Диалог-2011». Вып. 10 (17). М.: Изд-во РГГУ, 2011. С. 510-522.

10. Webb, G. et al. Not So Naive Bayes: Aggregating One-Dependence Estimators // Machine Learning. 2005. 58. Pp. 5-24.

11. Hatzivassiloglou, V., McKeown, K. Predicting the Semantic Orientation of Adjectives // Proc. of the 35th Annual Meeting of ACL, Madrid. 1997. Pp. 174-181.

12. Маннинг К., Рагхаван П., Шютце Х. Введение в информационный поиск. М.: Вильямс, 2011. 520 с.

REFERENCES

1. Carenini, G., et al. Extracting Knowledge from Evaluative Text. *Proceedings of the 3rd International Conference on Knowledge Capture*. 2005. Pp. 11-18.

2. Hu, M., Liu, B. Mining and Summarizing Customer Reviews. *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2004. Pp. 168-177.

3. Nasukawa, T., Yi J. Sentiment Analysis: Capturing Favorability Using Natural Language Processing. *Proceedings of the 2nd International Conference on Knowledge Capture*. Florida. 2003. Pp. 70-77.

4. Pang, B., Lee L. A Sentimental Education: Sentiment Analysis Using Subjectivity Summarisation Based on Minimum Cuts. *Proceedings of the ACL*. 2004. Pp. 271-278.

5. Turney, P. Thumbs Up or Thumbs Down?: Semantic Orientation Applied to Unsupervised Classification of Reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. 2002. Pp. 417-424.

6. Ermakov, S.A., Ermakova, L.M. Overview of Sentiment Analysis Methods. *Vestnik Permskogo universiteta — Perm University Herald*. 2012. Issue. 1(19). Pp. 85-89. (in Russian).

7. Lukashevich, N.V., Chetverkin, I.I. Retrieval and Application of Sentiment Lexicon in the Context of Reviews Classifying into Three Classes. *Vychislitel'nye metody i programmirovanie — Computational Methods and Programming*. 2011. V. 12. Pp. 73-81. (in Russian).

8. Orobinskaja, E.A., Kochueva, Z.A. Text Mining Techniques: Review of Methods and Tasks of Content Processing. *Vestnik Hersonskogo nacional'nogo tehnicheskogo universiteta — Herson National Technical University Herald*. 2010. № 2 (38). Pp. 348-353. (in Russian).

9. Pazel'skaja, A.G., Solov'ev, A.N. Method of Emotion Determination in Russian Texts. *Komp'juternaja lingvistika i intellektual'nye tehnologii: «Dialog-2011» — Computational Linguistics and Intellectual Technologies*. 2011. Issue. 10 (17). Pp. 510-522. (in Russian).

10. Webb, G. et al. Not So Naive Bayes: Aggregating One-Dependence Estimators. *Machine Learning*. 2005. 58. Pp. 5-24.

11. Hatzivassiloglou, V., McKeown, K. Predicting the Semantic Orientation of Adjectives. *Proc. of the 35th Annual Meeting of ACL*. Madrid. 1997. Pp. 174-181.

12. Manning, Ch., Raghavan, P., Schütze, H. *Vvedenie v informacionnyj poisk* [Introduction to Information Retrieval]. Moscow, 2011. 520 p. (in Russian).

Авторы публикации

Брунова Елена Георгиевна — заведующая кафедрой иностранных языков и межкультурной профессиональной коммуникации естественнонаучных направлений Института математики и компьютерных наук Тюменского государственного университета, доктор филологических наук, профессор

Бидуля Юлия Владимировна — кандидат филологических наук, доцент кафедры информационных систем Института математики и компьютерных наук Тюменского государственного университета

Authors of the publication

Elena G. Brunova — Dr. Sci. (Philol.), Professor, Head of the Department of Foreign Languages and Cross-Cultural Communication in Science, Institute of Mathematics and Computer Sciences, Tyumen State University

Yuliya V. Bidulya — Cand. Sci. (Philol.), Associate Professor, Department of Information Systems, Institute of Mathematics and Computer Sciences, Tyumen State University